ISSN 2013-9004 (digital); ISSN 0210-2862 (paper)
http://dx.doi.org/10.5565/rev/papers.2091
Papers 2014, 99/4
595-629

Reasons and biological causes. Some reflections on Boudon's Theory of Ordinary Rationality*

Francisco José León Medina
Universitat de Girona
francisco.leon@udg.edu

Received: 22-04-2014
Accepted: 28-05-2014

Abstract

In his effort to provide sociology with a theory of behavior for the microfoundation of social phenomena, Raymond Boudon searched for a theory that could be presented as *general* (i.e., a theory that, given its strength, can be used "by default" in explanations). In this search, Boudon disregarded biological causes and stated that his Theory of Ordinary Rationality was the best choice, since it offers final explanations: when a behavior is explained as a result of beliefs that are grounded on good reasons, we are offering a black-box-free explanation. In this paper, I shall state that there are serious problems in the arguments that Boudon used to dismiss the explanatory strategy based on "biological causes". Secondly, I shall point out that some recent findings of several behavioral sciences constitute a radical questioning of the value of his Theory of Ordinary Rationality, as well as a positive revaluation of Evolutionary Psychology. In light of these findings, we can state that on many occasions, either reasons are systematically biased by biological causes, or these causes cause behavior, thus reasons are mere rationalizations. Therefore, neither the reason-based explanatory strategy nor the biological causes-based explanatory strategy can be used "by default". Given the current state of our knowledge, Evolutionary Psychology cannot stand as a *general* theory of behavior but is better placed to do the job in the future: it will depend on its ability to build models that integrate reasons and biological causes.

Keywords: rationality; rationalism; reasons; evolutionary psychology; adapted mechanisms; biological causes; Raymond Boudon.

* This work has benefited from a MINECO R&D project grant (reference number CSO2012-31401) and from a CONSOLIDER-INGENIO 2010 project grant (reference number CSD 2010-00034).

Resumen. *Razones y causas biológicas. Algunas reflexiones sobre la teoría de la racionalidad ordinaria de Boudon.*

En su esfuerzo por proveer a la sociología de una teoría del comportamiento con la que microfundamentar los fenómenos sociales, Raymond Boudon buscó una teoría que pudiese presentarse como *general* (es decir, una teoría que, dada su fortaleza, pudiera usarse «por defecto» en las explicaciones). En esta búsqueda, Boudon desconsideró las causas biológicas y defendió que su teoría de la racionalidad ordinaria era la mejor elección, pues ofrecía explicaciones finales: cuando una conducta se explica como resultado de creencias que están bien fundamentadas en buenas razones, estamos ofreciendo una explicación libre de «cajas negras». En este artículo, sostendré que existen serios problemas en los argumentos que usó Boudon para descartar la estrategia explicativa basada en las causas biológicas. En segundo lugar, señalaré que algunos hallazgos recientes de varias ciencias del comportamiento suponen un cuestionamiento radical del valor de su teoría de la racionalidad ordinaria así como una reevaluación positiva de la psicología evolucionista. A la luz de estos hallazgos, podemos sostener que, en muchas ocasiones, o bien las razones están sistemáticamente sesgadas por causas biológicas, o bien esas causas causan directamente la conducta, por lo que entonces las razones son meras racionalizaciones. Por lo tanto, ni la estrategia explicativa basada en razones ni la basada en causas biológicas pueden usarse «por defecto». Dado el estado actual de nuestro conocimiento, la psicología evolucionista no puede proponerse como una teoría *general* del comportamiento, pero está mejor situada para hacerlo en el futuro: dependerá de su habilidad para construir modelos que integren razones y causas biológicas.

Palabras clave: racionalidad; razones; racionalismo; psicología evolucionista; mecanismos adaptativos; causas biológicas; Raymond Boudon.

Summary

1. Introduction

Almost a dozen decades after the publication of Durkheim's *Rules of Sociological Method*, most sociologists are still stuck in the foundational error of the discipline: the idea that social facts are *sui generis* and hence irreducible to lower-level facts. Armed with an argument that legitimates a frequently proud ignorance of the developments of the rest of the sciences, most sociologists keep themselves outside of the project that places the hope for a renovation and a better future for the discipline in the microfoundation of social phenomena. Raymond Boudon has certainly been one of the exceptions in this gloomy picture, and his effort to provide sociology with a theory of behavior for the microfoundation of its explanations shined as few others did.

With this project in mind, Boudon addressed the evaluation of the available theories in search for some theory that could be presented as a *general* theory of social behavior. It should be pointed out that the exercise consisting in evaluating theories comparatively is not only a legitimate exercise, but a desirable one. In a normal science, theories do not peacefully coexist, isolated from one another in their ivory towers. Weaker theories surely would interpret the comparative assessment as an aggressive attack on a supposed desirable diversity, but respect for weak theories is not what made science knowledge progress (quite the opposite).

In the case of Boudon, this comparative evaluation is also made with the purpose of identifying a general theory of behavior. Even though Boudon did not explicitly state what he meant by "general theory", one can infer from his writings that he was referring to a theory that, given its strength, can be used "by default" in explanations. For example, if rational choice theory is considered a general theory, then it would make sense to establish the methodological principle of rationality: if a behavior is to be explained, the reasonable thing to do is to start with rationality-based explanations (1998b: 174). Thus, using a type of explanation "by default" would not be a guarantee of success, since other causal forces could be operating, but it would be the least bad of all methodological strategies.

That being said, what is really questionable is the way Boudon makes this comparative assessment. In this paper, I shall try to defend the following arguments. First, I shall state that there are serious problems in the arguments that Boudon uses to dismiss the explanatory strategy based on what he called "biological causes" of behavior. Secondly, I shall point out that some recent findings of several behavioral sciences constitute a radical questioning of the value of his Theory of Ordinary Rationality (TOR from now on), and generally of the rationalist or utilitarian-cognitivist paradigm, as well as a positive revaluation of the Evolutionary Theory of social behavior or Evolutionary Psychology (EP from now on). And lastly, I shall maintain that, in the light of these findings and the current state of our knowledge, neither the reason-based explanatory strategy nor the biological causes-based explanatory strategy can be proposed as a "by default" strategy in the explanation of social behavior, although EP is better placed to do the job. In this sense, the fundamental challenge of the microfoundation strategy of social phenomena will be its ability to generate useful knowledge and criteria to determine which of these two strategies or combination of both is appropriate in each case.

1.1. Evaluating theories of social behavior

In his discussion on the sociology that really matters (2002), Boudon presented a set of criteria for judging what a scientific theory is and how to judge its strength. It seems clear that he used these criteria in his comparative evaluation of different theories of behavior (2006, 2007, 2009). These criteria are

uncontroversial, but could be more accurately developed. This is why we shall present them together with our proposed extension.

First, Boudon proposed the criterion of logical consistency of the propositions that form a theory (2002: 373): a theory cannot house inconsistent propositions. The criterion is indisputable, but it would also be necessary to point out that logical consistency is not a purely internal problem of the theory, so that we could extend the list of Boudon's criteria adding the following one: the propositions of a theory must not conflict with principles and findings that are already well established in other sciences.

Second, Boudon pointed out the criterion of the acceptability of explanatory propositions (2002: 373). There are several reasons for the acceptability of a proposition, but Boudon especially stressed two: propositions must have empirical support and must not include obscure concepts. At least for realists, it is sensible to consider that empirical support must also be on the basis of the principles and assumptions of the theory, at least if we agree that the purpose of the theory is to explain and not only to predict. Therefore, the empirical support for the principles and assumptions of the theory is a criterion that could also be added to the list of those proposed by Boudon.

Thirdly, during his comparative evaluation of theories of behavior, Boudon also used the criterion of the explanatory scope of the theory (2006, 2007): theories can compete in their ability to integrate dispersed empirical results, so we expect that a general theory of behavior would not leave some relevant phenomena unexplained. However, the explanatory scope of a theory is not only an external problem: the process of logical inference from the postulates of the theory must be shown to be fertile in its ability to generate new testable predictions. A fertile theory in this sense is a theory with greater explanatory scope, so that fertility could also be added to the criteria proposed by Boudon.

Therefore, our extended version of the criteria used by Boudon leads us to the following six criteria: a) internal logical consistency, b) external logical consistency (with other disciplines), c) acceptability of its propositions, d) acceptability of its assumptions, e) explanatory scope, and f) fertility. Throughout the paper, we shall use these six criteria for judging Boudon's comparative evaluation of different theories of behavior and for revaluating it in the light of the new findings of several behavioral sciences.

2. Reasons and causes in Boudon's writings

2.1. The Theory of Ordinary Rationality

Boudon was, along with Jon Elster, one of the authors who most acutely addressed the limitations of Rational Choice Theory (RCT from now on). While it is true that replacing an explanation of behavior based on *cultural forces* (such as socialization) for one based on RCT causes an advance in knowledge (see, e.g., Boudon, 2006), several theoretical and empirical reasons led Boudon to argue that RCT could not aspire to be a general theory of social behavior (1998a,

2006). From a theoretical point of view, it seems clear that a) not all actions are instrumental, b) not all instrumental actions are guided by the criterion of utility maximization, and c) RCT does not have much to say about the beliefs, values and objectives on which the action is based. In this sense, RCT has a problem with the fourth of our evaluation criteria (*acceptability of the assumptions*). From an empirical point of view, RCT is incapable of explaining phenomena such as the paradox of voting or several behaviors that are usually observed in the experiments of behavioral economics. RCT, therefore, also has a problem with the fifth evaluation criterion (*the explanatory scope*).

Following the lead of Simon (1982), Boudon grounded his theory in the notion of *subjective rationality*. In objective rationality, the reasons guiding the subject are objectively valid, and therefore there is no mystery as to why the subject perceives them as good reasons. In *subjective rationality*, reasons are not objectively grounded, yet they are perceived as good (Boudon, 1989). Given the demanding conditions to assert that a reason is objectively grounded, it was not difficult to conclude that in most situations the reasons that move us are subjective, and that this is something that our theory of behavior must take into account. In that way, Boudon connects with the Weberian tradition, according to which beliefs and actions can be objectively unfounded and still be understandable (i.e., explainable as a result of reasons perceived as good). Thus, he took on the challenge of building a theory of behavior based on an extended notion of rationality, not on rationality in the strict sense, but rather on reasonableness. Throughout the years, he gave different names to that theory: Cognitivist Model of Rationality (1996), Rational Model in the Broad Sense (2000), Theory of Ordinary Rationality (2009).

TOR seeks to explain the adherence of individuals to "goals, values and representations" (Boudon, 2009: 58). According to this theory, individuals adhere to a goal, value or representation when they perceive it as the consequence of a set of reasons composed of acceptable and compatible elements, and provided that there is no other preferable alternative set of reasons. Boudon called this the *cognitive equilibrium principle* (2012). Thus, the system of reasons cause the acceptance (and the strength of the acceptance) of the individual to that goal, value or representation.

According to Boudon, TOR keeps one of the main advantages of RCT: its final character. When an action is explained as the result of goals, values or representations that are based on good reasons we are also providing a black-box-free explanation. Thus, TOR is presented as a theory that posits that positive or representational beliefs (of the type *X is true*) and normative beliefs (of the type *X is good, fair, legitimate...*) are rational, that is, they are grounded on good reasons (Boudon, 2009). Actions motivated by these beliefs are then rational too.

Especially in *The art of self-persuasion* (1990), but also in other studies (e.g., 1989), and following the lead of Weber, the more empirical Durkheim and some argumentative line of Simmel, Boudon showed that false beliefs, both individual and collective, are also understandable, that is to say, they are the

result of a coherent set of reasons that the individual perceives as acceptable. Undoubtedly, one of the great contributions of Boudon was to note that when we face the existence of false beliefs as scientists, the dead easy recourse to an irrational-based explanation, such as when we "explain" behavior as a result of socialization, offers less satisfactory explanations than those offered by TOR (see, for example, 1989). In fact, Boudon presents this dead easy recourse as part of a "spontaneous sociology" opposed to a scientific sociology (1990: 18). As we shall see, it is questionable that the same argument can be applied to the biological causes-based explanatory strategies.

2.2. Social causes: a demolishing critique

When Boudon developed a typology of theories of behavior (2006), as well as when he developed a typology of theories of values (2001), he classified evolutionary and cultural explanations in the same category. Although both explanations have seemingly little to do with each other, in fact the classification makes sense, since both explanations maintain that mental states can have causes that are unnoticed by the individual, rather than pointing to reasons as TRO does. In the words of Boudon, both theories understand mental states as caused, not as grounded (2001: 32).

Boudon addressed his criticism of *social causes* in his assessment of cultural forces-based explanations. We define *social causes* as those social structures that supposedly shape the individual mental states (Lizón, 2010). Boudon's critique of the explanatory power of these causes is demolishing (see, for example, 1990, 2006). The idea that individual beliefs are mere reflections of collective beliefs, manifested in the individual through socialization, is a surprisingly popular pseudo-explanation, but its fragility and inconsistency is obvious if analyzed in minimal detail. For Boudon, cultural explanations, such as the explanation of LévyBruhl of magical beliefs as a result of a "primitive mentality", are based on cumbersome psychological hypotheses and on *ad hoc* built concepts leading to tautological explanations (the "primitives" confuse verbal associations with causal relationships because they have a primitive mentality, and that mentality consists of a tendency to confuse verbal associations with causal relationships).

Fascinated by the huge diversity of human cultural forms (and to a large extent overestimating it), twentieth-century social science felt compelled to explain certain practices. With the weakness of his conceptual apparatus, the alternative of appealing to the effects of socialization used to generate the false impression of having solved the puzzle. To the question "why do members of culture *x* do *y*?" one could answer "because they have been socialized to do *y*". This apparently deep proposition says practically nothing. The expression "have been socialized to do *y*" is equivalent to the expression "have learned that in culture *x* people do *y*", so that by replacing this proposition for the original, the initial proposal states that "members of culture *x* do *y* because they have learned that members of the culture *x* do *y*". Obviously, the original

question still stands, and rather now the question is twofold: on the one hand, we might wonder about the origin of this cultural practice, and on the other hand, we might wonder about the reasons for the individual's adherence to it (since the mere transmission from one generation to another does not explain its acceptance by the receiving generation: one would need reasons to maintain a belief learned through socialization – Boudon, 2001: 5-6). The problem of circularity in socialization-based explanations points at a problem of the cultural forces theory with the third of our evaluation criteria (*acceptability of the propositions*). This acceptability is further compromised by the constant presence of ill-defined, ambiguous and obscure concepts such as *habitus*, primitive mentality, etc. (Boudon, 2006).

But the problems do not end there. Boudon also notes that these explanations have problems with the first criterion (*internal logical consistency*). For example, in the case of the prevalence of the rule of unanimity in the collective decisions of rural Vietnamese societies, the theory of cultural forces states that in traditional rural areas the individual is subject to the group, and only a unanimous decision can be regarded as legitimized by the group. However, it is not difficult to see that the rule of unanimity is precisely the rule that gives more power to the individual over the group, as unanimity is synonymous with veto power (2006: 153).

Furthermore, cultural theories are weak in generating empirically testable theoretical predictions (sixth criterion). In that sense, Boudon appeals to the uncertainty about the effects of socialization. On the one hand, we know that socialization is not always successful, but the theory does not provide elements to predict when it will not be (e.g., Boudon points to Weber's analysis of the sudden conversion of the Roman civil servants and military officials to Monotheism – 2006: 181). On the other hand, it is known that socialization can have opposite effects: an alcoholic father can either become a role model for his child, driving him to alcoholism, or become a negative model to avoid, leading him to abstemious behavior. Thus, the theory loses its scientific character, since *a posteriori* it is always able to interpret any observable effect as consistent with the proposed cause (or in other words, the theory does not provide tools for its refutation).

This does not mean that there are not *social causes*: socialization obviously exists and has an influence on our beliefs and behavior. But, since cultural theory is affected by so many problems, Boudon rightly concluded that it could not aspire to become a general theory of social behavior (1998a, 2006).

3. Boudon's disdain for the biological causes

In the context of theories of behavior, we define *biological causes* as those neurophysiological processes with a genetically conditioned structure and function, which are activated and modulated by different (material or social) environmental inputs, operate outside the consciousness of the individual,

and have a direct or indirect systematic influence on behavior. In fact, the distinction between *reasons* and *biological causes* is very problematic. Are not reasons neurophysiological processes? Reasons *are* a biological phenomenon. However, we will reserve the term *reasons* to refer to conscious mental representations consisting of arguments for or against a proposition or a set of propositions. It is obvious that these mental representations are the emergent effect of neurophysiological processes, but this is relatively unimportant in the context of this paper. The concept of *biological causes* is reserved here to refer to nonconscious processes operating either on behavior or on mental representations that govern behavior.

Given that there are no doubts that biological causes do exist, the debate for social scientists has focused on their relative importance and on the role they must play in the theory of social behavior. Boudon, and analytical sociologists overall, have played a fundamental role in the erosion of the false belief that social facts are *sui generis*, stating that the microfoundation explanatory strategy should be the proper sociological explanatory strategy, which certainly involves questioning the boundaries of sociology and psychology. But, with some exceptions (Lizón, 2010), the first generation of analytical sociologists stated their preference for intentional explanations of social action, and therefore felt some vertigo when facing the final consequences of the openness to explanations based on biological causes.

As already mentioned, the Boudonian approach to rationality is based on Simon's distinction between objective and subjective rationality (1982). In its definition of subjective rationality, Simon refers to an action that is appropriate to the achievement of given goals within the limits imposed by exogenous (environmental characteristics) and endogenous (characteristics of the organisms) conditions or constraints. Boudon does not seem to develop all the implications of this crucial point of Simon. In his concept of cognitive rationality, he accepts that there are exogenous constraints on what we consider good reasons: different social contexts can result in different sets of reasons being more easily evoked and accepted (2003: 16, 2009: 63). However, he hesitated at how endogenous constraints should enter the model. Faced with this challenge, Boudon moved between questioning the concepts employed in the theory of biological causes and questioning the role to be reserved to these causes.

First, Boudon noted that the theory of "biological forces" had problems with the third of the evaluation criteria (*acceptability of the propositions*). In particular, he pointed out that concepts such as those of bias or risk aversion are just descriptive and *ad hoc* concepts leading to circular explanations. Thus, for example, appealing to the availability heuristic to explain an overestimation of probabilities would provide a tautological explanation: an individual tends to overestimate the probability of an event when it is easily accessible (with known, experienced or easy to remember examples) because he has a tendency to overestimate the probability of an event when it is easily accessible. Now, even though this critique seems solid, it faces a major objection. Consider,

for example, the behavior of some animals (including humans) consisting of preparing the nest (or equivalent) during pregnancy. Biologists have explained this behavior as a consequence of a nesting instinct. If the nesting instinct is defined as a tendency to prepare the nest during pregnancy, does this mean that biologists are offering a circular explanation? Obviously not, provided that it is justifiable to state that the instinct conceivably exists. But this is precisely what evolutionary theory does: to argue that biases (for example) result from a predisposition resulting from an adaptive process. Boudon himself admitted that "It [the notion of bias] could cease to be a mere word if it could be shown that biological evolution, say, has produced a wiring of the brain explaining the bias" (2006:159). But here Boudon confuses the biological and cognitive levels. It is not strictly necessary to identify the neural basis of a cognitive trait to defend its adaptive nature. In fact, the ways to support the plausibility of the existence of a "natural" predisposition are diverse: the neurophysiological basis is one, but also its ontogenetically early appearance, its presence in other primates, its universality in the human species, its functionality for certain adaptive challenges, etc. To the extent that the empirical findings of cognitive psychology have been restated by EP as evidence of adaptive cognitive-behavioral programs (in the next section we will see several examples of this), Boudon's objection is neutralized and the explanatory potential of EP reinforced.

As we said before, Boudon also believes that the acceptability of the propositions of a theory depends on its empirical evidence problems. In that regard, he noted that, in general, evolutionary explanations suggest a phylogenetic conjecture that it is hard to prove (1996: 130). This is a classic critique of EP, but it is generally based on the ignorance of the real heuristic discovery process that this theory uses (see, for example, Machery, forthcoming; Schmitt and Pilcher, 2004).

Second, Boudon attacked the theory of "biological forces" referring to its alleged problems with the first criterion of evaluation (*internal logical consistency*). Thus, for example, he pointed out an alleged contradiction between natural selection and the existence of cognitive biases that systematically lead us to forecast errors (1996: 130). Boudon commits a fundamental error here: either he considers that adaptive designs of the past necessarily have to be adaptive in the present, or that adaptive designs of a context cannot be activated with harmful effects in other contexts. Both possibilities are wrong. As Gigerenzer could see, cognitive biases identified from the work of Kahneman and Tversky are adaptations that take most of the world's regularities. But the world in which our brain evolved into its current form was the Paleolithic, not that of our societies.

And thirdly, Boudon highlighted some problems with the *explanatory scope* of the theory of "biological forces". For example, he noted that this theory cannot explain why in some experiments the answers are so sensitive to changes in the problem formulation (1996: 130). In fact, a main assumption of EP and cognitive psychology is that adaptations are extremely sensitive to contextual cues and, consequently, different cues can trigger very different behavioral

programs (Tversky and Kahneman, 1981). For EP, behavior is extremely context-dependent.

As mentioned, in spite of all these criticisms, Boudon also addressed the question of the role that should be reserved for biological factors in explaining social behavior. His position here is far from clear, but in general, he noted that those forces must play *some* role. Referring to neuroscience, for example, he stated that "it can effectively contribute to the explanation of phenomena of interest to social science" (2009: 112). Surprisingly, the statement was not accompanied by any effort to integrate these contributions into his TOR, probably because his idea of the contribution of these disciplines was wrong. In his text *La racionalidad en las ciencias sociales* (2009) he presented two examples: that of an individual whose optimism was the result of a calcification in his amygdala and that of the acceptance of unfair proposals in an ultimatum game as a result of the neutralization of the activity of the dorsolateral frontal cortex. What is implicit in the text is that neuroscience could explain exotic or strange behaviors that result from the peculiarities of a special brain or from the sectorial paralysis of its normal activity. Thus, Boudon is omitting the real contribution of neuroscience: revealing how everyday actions and decisions of people with normal brains are related to processes that are beyond our consciousness.

Boudon questioned that *biological causes* could be the basis of a general theory of behavior (1996, 2001, 2006, 2007). We shall discuss if he was right or wrong in the last section. The problem is that, based on the errors and the unfounded criticisms presented above, he also ruled out *biological causes* as a key element in the explanation of social behavior. Below, we shall address the implications of this positioning.

3.1. A black box inside the black box

Boudon repeatedly noted that explanations based on the reference to psychological forces (such as those contained in the concepts of *bias* or *frame*) or biological forces (such as those characteristic of sociobiology) are problematic (1998a: 820; 2003: 3). The main reason was that their inclusion was supposed to necessarily derail the final nature of explanations based on ordinary rationality (2009: 116-117). Incorporating notions as *bias* or *module* a black box appeared where initially there was a final explanation, since these concepts relate to elements that are not self-explanatory and whose origin is unknown. Moreover, in many cases such terms would refer to confusing concepts, and Boudon even stated that they were "mere conjectures" (2006: 151) or "mere words" (1998a: 820).

However, Boudon also recognized that various "forces" that are not reasons can affect our beliefs and actions. He did so, for example, when he recognized that a belief can be explained by unconscious mechanisms such as adaptive preferences, or forces as passions (for example, he noted that jealousy can cause the belief in infidelity despite the absence of good reasons to support this

belief) (1990: 4). In *La racionalidad en las ciencias sociales* (2009), he stated that "reasons grounding a belief [...] can be biased under the action of various mechanisms. But adherence to a belief is always the effect of reasons" (2009: 87). Boudon, therefore, was open to recognizing the existence of systematic (biological or not) biases in the reasons grounding our beliefs and actions. Moreover, he acknowledged that there is not a general criterion on the strength of a set of reasons, and all that can be said is that we accept a set when we cannot imagine a better alternative set (2003: 16-17), but this also raises doubts about the ultimate causes of that strength.

The obvious problem for Boudon's position is that, if there are systematic but not studied biases in the persuasion power of a reason, TOR would be assuming a black box in its explanation. How can we assert the existence of systematic biases (on the strength and direction of a reason) the explanation of whose functioning we choose not to consider in our theory, and state at the same time that the main virtue of that theory is the absence of black boxes in its explanations? The inevitable conclusion is that progress in understanding *biological causes* is showing the existence of a black box inside the black box: Boudonian identification of the set of reasons grounding a representation can open the black box that cultural theories and behaviorism blithely assumed, but this is often insufficient to ensure the final character of the explanation. In short, this is not about explanation of behavior being necessarily based either on reasons or on biological causes, but about these two causal forces operating in some combination that 21st-century behavioral science will have to unravel.[1] As we shall see in the next section, EP is offering an evolutionary explanation of our set of adapted cognitive mechanisms, thus letting us go one step further in the process of microfoundation of social phenomena.

4. The challenges of behavioral sciences

As we shall try to argue in this section, the paths of behavioral sciences in the 21st century are inevitably leading us to question the value of TOR as a general theory of behavior. The illusion that a general theory of behavior could do without the so-called "biological forces", despite the recognition of their systematic influence on mental representations and behavior, seems to have its days numbered: the biochemical can no longer be left out of the analysis of the psychosocial. In considering the role of "biological forces", some empirical findings from very different disciplines and research areas are revolutionizing our conceptions of how we perceive, reason, decide, make moral judgments, enjoy the aesthetic, etc. TOR is not only unable to reconcile these results with

1. The distinction between proximate and ultimate causes is a useful analytical tool, but behavioral science cannot settle for an appeal to a supposed difference between two alternative "levels of explanation": an interpretive framework identifying the articulation of both causal forces is required.

the theory, but contradicts them, and is therefore seeing the acceptability of its assumptions and propositions very threatened. Facing TOR, EP is not only providing an interpretive framework to give coherence to all these findings, but is often serving as a generating matrix thereof.

To illustrate this argument, we shall conduct a comparative evaluation of TOR and EP in three different sections. First, we shall address how our understanding of how we process information (how we perceive, think, etc.) has been revolutionized. Second, we shall present some examples of how our explanation of the decision process (especially in economic issues) has been modified. And finally, we shall focus on how our ideas about how we make moral judgments have been challenged. These three fields (information processing, decisions, moral judgments) will serve to illustrate the battle of the two theories. For reasons of space, we cannot address other examples, but this analysis could be extended much further. For example, to assess the ways in which we assess the sex appeal of potential mates and how we choose mates, how we shape our magical and religious beliefs, how we form social hierarchies, etc.

4.1. How we process information

During the 20th century, and under the influence of behaviorism, it was believed that the human being was a blank slate that, under the right stimulus program, could end up showing (within the obvious biological limits) any belief, preference, skill, or behavior. The human being would only bring with itself a general capacity for learning and abstract reasoning, and the rest would be a result of the inputs coming from the social environment.

Despite their differences, both cultural theories and rationality-based theories are part of the Standard Model of Social Sciences (as Tooby and Cosmides called it – 1992). Faithful to this assumption on the human mind, the adherence to beliefs has been understood in TOR as a result of a general reasoning ability: subjects would arrive at different beliefs because they are grounded on different sets of information, but everybody employs the same general intelligence, the same rules of abstract inference. That is how Boudon explains, for example, the "rationality" of the primitive's magical beliefs (see, for example, 1989: 180; 2009: 69), but also scientific beliefs, normative beliefs, and any other type of mental representation. One and the same system of information processing (a general intelligence) would be grounding our good reasons to lend (or not) money to a friend, to morally condemn someone else's behavior, or to judge a potential partner as desirable.

TOR, and in general all the theories based on rationality, are seeing this assumption challenged as a result of the confluence of several disciplines on the same approach: that the human mind, far from being a blank slate, is equipped with a set of psychological modules containing representations and content-specialized processes activated as a result of specific inputs from the environment and prefiguring automatic and predesigned responses (Cosmides and Tooby, 1994; Pinker, 2003; Tooby and Cosmides, 1992). Our mental

architecture has a domain-specific organization. The alleged indifference to the different stimuli is simply not true: the human mind processes different types of information differently, and that process involves many different cognitive-behavioral and partly instinctive, unintentional and nonconscious processes. This conclusion is reached, in fact, having to fight the ideological appeal and the apparent obviousness of the theory of the black slate, and accumulating evidence especially coming (but not only) from neuroscience and cognitive psychology. From a neurophysiological point of view, the evidence that our mind does not process all inputs with the same system comes largely from the study of the effects of neuronal injuries. Localized lesions affect specific functions without overall harm to the general cognitive ability of the individual. For example, there are individuals who maintain their ability to distinguish any two material objects but are unable to distinguish two human faces, two animals of different species, or two fruits. From a cognitive point of view, it has been shown, for instance, that different animals have an instinctive fear of specific predators despite not having seen them in their entire life, or even despite the species having been isolated from them for thousands of years (see, for example, Barrett, 2005). Unfortunately for culturalists and creationists, the human being is not different: a line of research (see, for example, Rackison and Derringer, 2008) has convincingly shown a predisposition to fear of snakes in humans and primates: snakes immediately catch our attention on visual complex arrays, it is easier to induce fear of snakes than induce fear to other objects, and it is more difficult to reverse that fear than the fear of other objects. And beyond this anecdotal and irrelevant predisposition from a sociological point of view, much more relevant evidence for social analysis is being accumulated: our economic choices, our moral judgments and our preferences in mate selection are shaped by specific modules containing inherited predispositions, but the list goes on almost indefinitely. To the dismay of advocates of socialization as a demiurge, a few minutes after birth babies follow stimuli similar to human faces more frequently than other stimuli, and show a difference between men and women, the former showing more interest in mechanical objects and the latter in human faces (Connellan et al., 2000); at two days of life they show a preference for their native language (Moon et al. 1993); at two months they make real social smiles (even if they are blind); at three months they already "know" some basic laws of physics (Baillargeon, 1987; Spelke, 1990); at 9 months they develop without instruction the so-called *joint attention* (using gaze direction of others to set their own) and they start to conceive others as intentional agents; at 18 months (and independently from encouragement and rewards) they show altruistic behavior (Tomasello, 2009), etc. And adults, despite the important role of socialization, also show partly automatic preferences and behaviors when they have to assess the beauty of a landscape (Orians and Heerwagen, 1992) or the artistic value of a work (Dutton, 2009), when they face the challenges of parenthood (see, for example, Gettler et al., 2011, 2012) or the threat of free-riders (Cosmides and Tooby, 1992, 2005), etc. All these predispositions would be nothing without the environment, but the information that comes from it would be nothing

more than an infinite chaos of bits of information if it were not for the existence of innate structuring structures in the human mind: without theories-formation mechanisms there cannot be learning.

Can a serious theory of behavior be grounded in the *tabula rasa* assumption when the idea of the equipotential mind is strongly discredited outside the culturalist stronghold which dominates the social sciences? Probably not. Boudon did not ignore these advances in our knowledge of the mind. In fact, on several occasions he acknowledged the falsity of the theory that states the human indifference to various stimuli, thereby accepting the idea that socialization works with and on innate predispositions (1997: 8, 2001: 77). But he did not develop this argument to its final consequences. From any point of view, TOR is part of the Standard Model of the Social Sciences.

The accumulation of empirical evidence from so many disciplines in favor of the idea of the adapted mind has led to a reversal of the burden of proof: it corresponds to the theory of the blank slate to explain how a general-purpose mind could evolve and produce the effects observed in the empirical work of those disciplines. And it does not seem to be having any success in this work. The consequences are devastating for TOR. A theory of behavior to be used as the basis for the microfoundation of social phenomena cannot contain statements that are inconsistent with those established in other sciences (as the *external logical consistency* criterion states) and cannot be grounded in questionable assumptions about human nature (according to our fourth criteria, the *acceptability of the assumptions*). The idea of not giving any role to biological processes in cognition conflicts with the uncontested evidence on the modularity of our mental architecture. Thus, its assumptions are deemed unrealistic and its propositions on the formation of beliefs are deemed inconsistent with propositions that are well established in other sciences.

Faced with this challenge, EP seems to be a much more solvent theory. From the evolutionary point of view, cognitive modules are conceived as adaptations: if there are specific systems to process specific types of information and containing pre-coded forms of reaction, it is because of their functionality in our past. These systems allowed us to effectively resolve recurring problems affecting our survival and reproduction during the Paleolithic. If, for example, we have a facial recognition module, it is because facial recognition had an adaptive function (to identify our people, remember past interactions, etc.). Our psychological architecture (the integrated set of instincts and general purpose mechanisms) comes from an evolutionary process. The mind is a product of the brain, and there is no special reason why the functional design of this organ has escaped the molding forces of natural selection. Conceiving modules as "designed" by natural selection to solve adaptive problems, EP avoids the problems that TOR had with the second and fourth evaluation criterion.

Moreover, EP has been very capable of empirically substantiating its propositions. Very contrary to the claims of some common critics, who point to a problem with the third of our criteria (*acceptability of the propositions*) because of the lack of empirical evidence, the proposition that a cognitive module is

an adaptation is usually not a mere *just-so story*: EP seeks confirmation of its hypothesis from a surprising variety of sources (see, for example, Schmitt and Pilcher, 2004).

The powerful theoretical framework of EP exceeds that of TOR also in the arena of other evaluation criteria. From the *explanatory scope* point of view, EP (especially because of its modular conception of the mind) appears capable of integrating empirical evidence that TOR could not accommodate in its approach, as we shall see in the following sections. For example, how could TOR explain the asymmetry – and the universality of the asymmetry – between women and men in their mate-choice preferences? EP has successfully done so (Buss et al., 1990). How could TOR explain that in certain social exchange situations evolutionary logic leads us to a logically incorrect but adaptive response? EP has successfully done so (Cosmides and Tooby, 1992, 2005). How could TOR explain why attractive men cooperate less in social exchange while attractiveness does not affect the probability of women cooperation? EP has successfully done so (Takahashi et al., 2006).

Also from the point of view of *fertility*, EP outperforms TOR. Conceiving cognitive modules as adaptations is proving to be a matrix for the generation of novel hypotheses about previously unknown psychological traits that end up becoming part of the explanation for some behaviors, something that TOR cannot claim to be able to do, since it is limited to explaining behavior *a posteriori*. If a psychological mechanism is conceived as an adaptation, that is to say, if we state that it has been shaped by natural selection to perform a specific function, then we can infer some attributes or components (usually referred to as "design features") that the mechanism is logically expected to have. For example, from *error management theory* it could be inferred that women (compared to men) underestimate the levels of romantic commitment that can be inferred from declarations of love. This design feature has been called *commitment skepticism bias*, and its existence has been confirmed by Buss (2000). As noted by Machery (forthcoming), grounding on assumptions about the adaptive nature of a psychological trait, EP is able to infer hypotheses about the existence of psychological capacities, the nature of the process, its development and some situational cues it uses.

4.2. How we make decisions

Simon's transition from objective rationality to subjective rationality was only the beginning of a long process that eventually led to the crisis of the theories that placed reasons as the only relevant causal force over our decisions. At first, the limited power of rationality was recognized, but advances in disciplines such as cognitive psychology and neuroscience took the argument further. It was not only about the existence of cognitive limits in the application of abstract reasoning to particular problems, but about the ubiquity of neurophysiological based cognitive-behavioral programs that, at least in part, precode ways of perceiving, evaluating and deciding.

In this section we present three examples of how the contributions of the behavioral sciences are putting TOR in check. Boudon faced several examples of these investigations (especially those from cognitive psychology – see, for example, 1990) and showed in a relatively acceptable way that they were reinterpretable from TOR. But this is far from being a proof of anything. If, say, the theory *A* is producing a series of results (a_1, a_2, a_3…a_n) that challenge the theory *B*, the exercise showing that a_1 and a_3 are reinterpretable from *B* faces the general problem of falsification: resistance to contrary evidence rather than favorable evidence strengthens the theory. If a_2 is still serving for the falsification of *B*, the exercise is futile. In addition, new empirical findings in favor of *A*'s interpretation of a_1 and a_3 are sufficient to get things back to place and for *A* to claim its superiority over *B*. As I shall try to argue in this section, this is exactly what is happening with TOR: new empirical findings of the behavioral sciences are providing evidence against TOR while EP seems to be in a better position to make sense of them.

Time inconsistency. One of the challenges to RCT that have arisen from these disciplines points to the socalled *time inconsistency*: imminent payments are more valued than future ones. If we get to pick a unit of a good in a month or two units in a month and one day, we will choose the two units waiting for a month and a day, but if we get to choose a unit today or two units tomorrow, many of us would choose a unit today. How could TOR explain time inconsistency? Is it possible to imagine any set of reasons according to which it is different to expect one day today than expecting one day in a month? In the absence of a declared set of reasons, one possibility would be to reconstruct the rationale underlying the decision, presenting behavior *as if* it were the result of this reasoning, but this strategy would lead us to a mere *just-so story*. Therefore, TOR faces in this case a problem with the fifth evaluation criterion (the *explanatory scope*), since it is unable to account for this phenomenon.

EP, however, offers a more satisfactory answer: our behavior is ecologically rational. What moves us is nothing but impatience, and impatience is an evolved mechanism that allows us to manage uncertainty. If there is a possibility that the payment will not be made, and we do not know the likelihood of that possibility, the passing of time can help us to assess it (Sozou, 1998), so that the adaptive response is to choose two units within a month and one day instead of one within a month, but if we have to choose between a unit today and two units tomorrow, we should ensure the profit since we do not know how likely it is that the payment will not be made. Whereas in our evolutionary past the chance that determines access to future resources was presumably high, developing an eager response to such decisions allowed us to profit from regular statistical patterns in the environment, thereby improving our fitness. Note that the decision is not grounded on reasons, but is caused by impatience, and here impatience is ecologically rational.

In the field of the explanation of time inconsistency, the superiority of EP does not only lie in its hypotheses generator matrix offering the explanation presented above, but also in the fact that a set of successfully tested predic-

tions whose results TOR would be unable to interpret have been inferred from that matrix. A first set of inferences are in the field of genetics. If time inconsistency is an adaptation, it necessarily has some support in our DNA. And indeed, there are some variants of genes that correlate with the tendency to show time-inconsistent preferences (Carpenter et al., 2011). Furthermore, by comparing twins a recent study has shown that "delay discounting" has a hereditary component (Anokhin et al., 2011).

Secondly, it would also be evidence for the consideration of time-inconsistency as an evolved cognitive-emotional program that there were a neurological system specifically involved in the phenomenon. Authors like Manuck et al. (2003) and Peters and Büchel (2011) have identified that system. Hariri et al. (2006), for example, have shown that the preference for instant but minor rather than larger and deferred rewards seems to be associated with the ventral striatum activity.

Third, in the study of the function of this neuronal system, the role of hormones is particularly relevant. Time-inconsistent preferences are activated as a result of environmental inputs, but these preferences should have a biochemical support. Here, Kayser et al. (2012) found evidence that dopamine reduces impulsivity in intertemporal choices, showing that hormones play a role in the structure of our temporal preferences.

Finally, another classic source of evidence of the adaptive nature of a psychological trait is primatology. One argument supporting the evolving nature of a trait (though in itself insufficient, like all others) is that it is not unique to humans but shared with our closest relatives. In this regard, it has been experimentally shown that this bias is not unique to humans, although it acquires specific features in each species. Non-human primates are also affected by it (e.g., rhesus monkeys – Hwang et al., 2009).

Therefore, we have evidence that suggests that genes play a role in this type of preferences, that there is a neuronal system involved in the phenomenon, that hormones play a role in this system, and that the trait is already present in other primates. In light of these results, the thesis that time inconsistency can be explained without reference to "biological forces" is untenable. Thus, with regard to time inconsistency, EP outperforms TOR (1) in *fertility*, as it is able to make new predictions (in fact, Boudon never offered evidence of TOR's fertility); (2) *acceptability of the empirical propositions*, as it is able to successfully test those predictions; (3) in *explanatory scope*, as it is able to integrate all these empirical results into a theoretical framework; and (4) to the extent that these results come from different disciplines, also in *external logic consistency*.

Loss aversion. A second example of empirical results that are better resolved by EP than TOR is loss aversion (Kahneman and Tversky, 1984), that is, the preference for avoiding losses rather than obtaining profits. Boudon considered this concept as merely descriptive, and thus the explanations of behavior as resulting therefrom, as circular. Again, this critique of the value of cognitive and evolutionary psychology due to its problems with the third criterion of evaluation (*acceptability of the propositions*) does not take into account some

of the most recent literature on the subject. Contrary to Boudon's assertions, the concept of loss aversion allows us to formulate very clear predictions about certain so far unknown biases, something that seriously questions that the concept has a merely descriptive character. For example, from the concept of loss aversion the existence of another phenomenon has been inferred: the so-called *endowment effect* (Kahneman et al., 1990). TOR does not have this fertility, and moreover it also has difficulties explaining these behaviors: how could ordinary rationality explain the gap between what we would be willing to pay for a product and what we would be willing to receive to sell it (Knetsch, 1989)?

In general, although we shall not go into detail, loss aversion bias has been explained by EP as an adaptation that seeks to maximize the number of offspring (see, for example, Levy, 2010) and initially seeks to maximize the acquisition of food resources (McDermott et al., 2008). If loss aversion is an adapted mechanism, it should be possible to find evidence of its genetic support, its biochemical basis, its neural organization and its relative continuity with non-human primates. And there is evidence in all these directions. Firstly, the variations of some genes are correlated with loss aversion. For example, there is some evidence that the serotonin transporter gene-linked polymorphic region (5-HTTLPR) polymorphism significantly influences performance in a Loss Aversion Task (He et al., 2010). In fact, beyond loss aversion, it seems that, in general, risk tolerance in financial decisions correlates with certain variants of some genes (Dreber et al., 2009; Kuhnen and Chiao, 2009; Zhong et al., 2009). By studying monozygotic and dizygotic twins, the heritability of economic risk preferences has been estimated to be 0.63 (Zyphur et al., 2009). Second, and consistent with the identified genes, the role of serotonin as a hormone that can lead to a reduction of loss aversion has been noted (Litt et al., 2006; Murphy et al., 2009). In general, risk behaviors in financial decisions are associated with the 2D:4D ratio, the ratio between the length of the second and fourth fingers; a ratio that depends on exposure to prenatal testosterone (Garbarino et al., 2011). Other studies have also indicated that risk behaviors in economic investments seem to have a nonlinear u-shaped relationship with endogenous testosterone levels (Stanton et al., 2011). Third, there is a difference in loss aversion between people with and without damage to certain parts of the brain (specifically in the amygdala, the orbitofrontal cortex and the insula, parts of the emotional brain) (De Martino et al., 2010; Shiv et al., 2005), suggesting that such areas perform some function in the phenomenon. And finally, it has been shown that some non-human primates show exactly the same loss aversion behavior (Brosnan et al., 2007; Chen et al., 2006).

Along with all this evidence of the evolutionary nature of this cognitive mechanism, the consideration of loss aversion as an adapted mechanism also allows us to elaborate new hypotheses on the cognitive field itself. EP, for example, has stated the domain-specific character of this mechanism, therefore predicting its variation in different contexts. As shown by Li et al. (2012), loss aversion is accentuated both in men and women when facing challenges in the

domain of self-protection, while it is erased for men facing challenges in the domain of mate selection, as inferred from EP.

In short, in relation to loss aversion, EP outperforms TOR insofar as the former is able to (1) provide an explanation of the phenomenon, (2) infer original predictions, (3) successfully test them, and (4) integrate all these different disciplines resulting in a single interpretive framework.

Social trust. The field of experimental economics, especially in connection with neuroeconomics, is also offering results that challenge the value of TOR as a general theory of behavior. Interestingly, Boudon referred to some of them as evidence of the limited character of RCT, but he could not note how far his theory was also challenged. To argue this point, we focus on a single example: experiments on trust.

In a trust game (Berg et al., 1995), an agent A (investor) receives an amount Y of money from the experimenter and has to send an amount X of money ($0 \leq X \leq Y$) to agent B (trustee). The investor keeps the amount that he does not send to the trustee. The experimenter multiplies X by a factor (for example, he triples it) so that the trustee has $3X$. The trustee must then freely decide how much (Z) he wants to return to the investor ($0 \leq Z \leq 3X$). So, the investor must decide whether to look for his own interest setting $X=0$, or to trust the trustee setting $X>0$.

As in many other experimental designs, RCT prediction is usually not fulfilled, as investors often transfer a positive amount to trustees. Can TOR explain why investors usually transfer a positive sum? Let's say a subject has decided to transfer 50% of his money to his opponent. Given the conditions of the experimental design (anonymity of the parties, the absence of reputation effects and shadow of the future, etc.), the subject probably has no option but to ground his decision on (1) his belief in the general level of people's trustworthiness, so he can treat his opponent under that criterion, and (2) his belief in the level of frustration he expects to experience if the trustee turns out to be untrustworthy. Both beliefs would be grounded on the past experience of the investor. The decision of how much to transfer would therefore be the result of a combination between trustworthiness expectations and betrayal aversion. For the subject, his decision to transfer 50% would be well grounded on beliefs for which he has good reasons (given the conditions of the experiment, there seems to be no other set of reasons that would justify the decision).

So far, TOR seems to be able to explain something that RCT cannot. However, experiments designed to test the role of oxytocin in these decisions can jeopardize TOR's explanation. In those designs all participants inhaled a product whose nature was unknown to them: half of them (the control) inhaled an innocuous substance and the other half (the treatment) inhaled a dose of oxytocin. The results indicate that there is a significant difference between the transfers of the two groups, being higher in the treatment (Baumgartner et al., 2008; Fehr et al., 2005; Kéri and Kiss, 2011; Kosfeld et al., 2005; Van Ijzendoorn and Bakermans-Kranenburg, 2012).

Now let's say that the same subject[2] who decided to transfer 50% of his money in the classic game (equivalent to the control condition in the oxytocin experiment) decides to transfer 70% to his opponent when subjected to the inhalation of the hormone. For the subject, his decision would, in both cases, be grounded on his belief in the general level of people's trustworthiness and his expectation about the level of frustration he would experience in the event of being betrayed. However, under the influence of oxytocin, the same evaluation leads to a different decision. Indeed, experiments show that the belief in other's trustworthiness is not altered between control and treatment, so that the most plausible hypothesis is that oxytocin affects betrayal aversion. When individuals are told to interact with a randomly acting machine, they do not modify their behavior despite oxytocin, which also reinforces the hypothesis of betrayal aversion (Elster, 2007).

Obviously, it is important to point out that what these experiments are showing is not that people become more trusting under the influence of oxytocin, but that *oxytocin levels affect trust levels*. It is not about 70% being a result of inhaling the hormone and 50% being the result of a decision process "free" of biochemical influences. This hormone is naturally produced in all of us,[3] and therefore it is logical to assume that what the treatment is doing is to increase its presence. The logical conclusion of the experiment is that our betrayal aversion in situations that require interpersonal trust is always influenced by oxytocin. As with the 50% transfer, TOR would explain the 70% transfer as the effect of good reasons: the individual believes that he will experience a low level of frustration in the event of being betrayed, which justifies a generous transfer. However, it seems clear that the explanation is inadequate if it simply refers to the belief system that the subject mentioned, since the good reasons grounding those beliefs are actually always biased by biochemical processes acting beyond the subject's awareness.

These experimental results turn the TOR explanation into a black-box explanation. If sets of reasons are not judged solely on their internal properties (consistency, acceptability, etc.) as stated by TOR, but are systematically affected by "forces" that we had not contemplated, the explanation of behavior as a result of a set of reason ceases to be a final explanation: a more fine-grain theory is needed.

Neuroscience and EP provide us the tools needed to open the black box inside the black box. The proposed mechanism is the following: oxytocin is a hormone that inhibits the amygdala, which is a center that is responsible for

2. Obviously, a subject cannot be exposed to the control *and* the treatment. Therefore, experiments analyze the difference in the average responses in the dependent variable between control and treatment groups. However, for the sake of clarity, we present the analysis of these results as if this problem of causal inference were not the case.
3. In reality, the degree of the oxytocin effect depends on its receptors, and the variability on those receptors depends on our genetic information. In the experimental deign, however, the random assignment of subjects to the control and the treatment group ensures an initial equivalence that allow us to test the average impact of a specific dose of the hormone.

emotional reactions, including fear, so that the hormone inhibits social fear, that is, aversion to being betrayed or exploited: it simply makes us more indifferent to the possibility that others do not honor our trusting behavior towards them. And what do the oxytocin levels which we are exposed to depend on? As could not be otherwise, they depend on the environment and genes. On one hand, experiences have an impact on the oxytocin level. Oxytocin levels lead us to a more trusting behavior, but at the same time, being a trustee also increases the levels of this hormone (Zak et al., 2005), leading us to behave as trustworthy. On the other hand, genes also play a role. An already identified gene encoding the protein OXTR, which is an oxytocin receptor, plays a crucial role. Depending on the allele of this gene, our oxytocin levels are higher or lower. Research correlating the three possible alleles of the gene (GG, AG, AA) with different social behaviors are a reality that only the most dogmatic sociologists can ignore (see, for example, Rodrigues et al., 2009; SaphireBernstein et al., 2011; Tabak et al., 2013; Tost et al., 2010; Walum et al., 2012). One of them has already provided evidence that individuals with the GG allele show a more trusting behavior than the rest in a trust game (Krueger et al., 2012).

While TOR has nothing to say about these biochemical processes that bias our perceptions and beliefs, EP offers an interpretative framework capable of integrating all these empirical results. From this theory, trust is interpreted as an adaptation whose function would be to enable cooperation and reciprocity where it is not possible to check the honesty of the other (Dunbar et al., 2007: 122), something that happens very often. Establishing relations of cooperation and reciprocity has obvious advantages for survival and reproduction, so the mechanism that makes this possible can be considered an adaptation. This evolutionary hypothesis allows us to interpret the role of genes and hormones in trust as biochemical processes resulting from natural selection in the environment of our ancestors. In fact, although some authors argue that it is not necessary to identify the biochemical processes that underpin what is proposed as a cognitive adaptation, the fact is that mind-brain unity suggests otherwise, so that the inability to detect a "biochemistry of social trust" would have been a setback for EP (and its presence is non-definitive but important evidence in its favor).

Moreover, the hypothesis is enhanced to the extent that several inferences derived from it have been successfully tested. For example, if trust has a role to play when it is not possible to check the honesty of the other, it is logical to expect that evolution has endowed us with a special sensitivity to scrutinize honesty cues. Since relatives are often trustworthy, phenotypic resemblance may act as one of these cues. DeBruine (2005) designed a trust game in which the investor was shown a picture of the trustee. In the control group a picture of a stranger was shown, while investors in the treatment were shown a photo that mixed the face of a stranger and the investor himself. Members of the treatment group were more likely to trust the trustee, confirming the inference obtained from the evolutionary theory of kin selection. When the phenotypic resemblance is not relevant or is simply absent, we seek other facial cues.

Using different brain imaging techniques, research such as that by Engell et al. (2007), Todorov et al. (2008), and Winston et al. (2002) show that there are specific areas of the brain that are activated to assess trustworthiness, contributing to the idea that trust is a hard-wired mechanism.

4.3. How we make moral judgments

A final example of the challenges that question the validity of TOR as a general theory of behavior comes from the science of the moral, and especially, from the confluence of research on moral philosophy and neuroscience.

Boudon postulated TOR as a theory that could also account for normative beliefs, and therefore moral convictions. For him, both positive and normative representations are always the result of a set of good reasons (2009). In *The moral sense* (1997) he explicitly addressed the question of the existence of innate and universal moral intuitions. In this text and in others, Boudon explicitly accepted the existence of an innate moral sense, but argued that its explanatory power of normative feelings was limited (1997: 9, 2001: 77). Boudon's objections were basically two. First, the theory of an innate moral sense would have difficulties explaining cultural variation. Second, although Boudon recognized that our assessments may be influenced by our human nature, the former generally cannot be deducted from the latter (1997: 9). Therefore, the criticism focuses on the fifth of the evaluative criteria (the *explanatory scope* of the theory). In the lines that follow I shall present some approaches to the science of morality that not only provide solid counterarguments to these objections, but also pose serious problems for TOR (problems that EP have no difficulty in solving).

That evolution has a role in the moral is something that Darwin himself had warned of (1871). However, the research that has shaped the evolutionary perspective of our moral sense is relatively recent. In an influential paper, Steven Pinker summarized this perspective (2008). For Pinker, our biological equipment incorporates a "moral switcher" that, when activated, leads to a special kind of reasoning (if you can call it that); a reasoning other than the one employed to determine if we like something, we are interested in it, etc. The rules that guide this mind-set are comparable to those of Chomsky's universal grammar: they are universal and they structure our moral intuitions in a way that goes unnoticed.

Brown's list of human universals (1991) includes a considerable number of aspects that can be considered typical of the moral: prohibitions such as incest, rape or violence, feelings such as shame, promotion of generous behavior and evil punishing, the distinction between good and evil, etc. Following this line, Haidt and colleagues noted that all cultures considered immoral things like hurting others, inequity, the lack of loyalty to the community, the lack of respect for authority and impurity (Haidt and Graham, 2007a; Haidt and Joseph, 2004). For them, these five principles are considered the ultimate psychological basis of all moral rules. Their universality is already evidence in

favor of its innate character, but so is that 1) all (except for purity) have some continuity in the behavior of other primates (de Waal, 1996), 2) an evolutionary history has been proposed for all of them (for a summary, see Haidt and Kesebir, 2010), and 3) some moral behavior and a distinction between moral rules and social conventions appear ontogenetically early (Tomasello, 2009; Turiel, 1983).

Boudon did not ignore this evidence, he simply pointed out (in his first objection) that the moral instinct could not account for the cultural diversity of moral conceptions. The error in this argument is clear: the moral instinct is universal, but apart from some behaviors that inevitably fall under the domain of morality (for example, rape or murder), others may moralize or amoralize depending on local, cultural processes (for example, in our societies tobacco consumption in public areas is no longer evaluated on pragmatic or instrumental criteria but it became moralized). Furthermore, the relative importance given to each of these principles varies between cultures and even between subcultures of a culture (Haidt and Graham, 2007b; Pinker, 2008). Boudon confused here the existence of a universal biological equipment to the defense of uniformity or universality of behavior; an argument that is clearly a *non sequitur*: at least part of our cultural conceptions are "evoked", that is, resulting from different inputs operating over a universal mental architecture.

In fact, EP has no problem in explaining cultural diversity as a result of universal predispositions (for a distinction between evoked and transmitted culture, see Tooby and Cosmides, 1992). For example, Boudon (1998) recalls the case of Madame de Sévigné, who in the seventeenth century wrote his daughter telling her how much she enjoyed attending a public execution. Today no one would admit to enjoying a public execution, says Boudon, and a "naturalistic" theory could not explain this cultural shift in what we consider moral. However, it is clear that in the case of capital punishment different moral intuitions conflict with each other: on the one hand, not hurting, and on the other, punishing the evil and being loyal to your people. The relative importance given to each of these principles varies between cultures, so understanding variation in moral judgments is not impossible from a "naturalistic" theory. What the "naturalistic theory" sustains (among many other things, as we shall discuss below) is that a) neither then nor now a normal person might consider moral the execution of an innocent, precious and prestigious member of the community (cultural variation has limits), and b) the effort to study the ways in which local variations in inputs that operate on mental architecture produce universal cultural diversity is or can be part of this theory.

The second objection that Boudon pointed out was that although our assessments may be influenced by our human nature, they cannot be deduced from it. The question, again, is whether these assessments can be understood without taking into account those influences. Our position is that they cannot. As we argued above, the recognition of systematic influences on human representations that we choose not to analyze constitutes an explicit waiver to developing a final theory, that is, an acceptance of a black box. Being universal

moral instincts the psychological foundation of all belief or moral judgment, no theory of moral beliefs can seriously do without their consideration. Boudon tried to do so, but to unravel the reasons that support normative judgments he was doomed to employ the fiction of the impartial spectator, who is able to put aside his interests and emotions and base his judgments only in the *common sense* (for example, 2009: 88). Boudon did not seem to wonder about the origins or ultimate foundation of that common sense, something that would have undoubtedly led him to research on the evolutionary basis of human morality. According to these investigations, Boudon's impartial spectator is equipped with innate predispositions to certain moral judgments. These predispositions also determine the type of moral reasoning that we do, and they do it in such a way that the principles of TOR are seriously threatened, as we discuss below.

Automaticity and rationalization. It is particularly important that the moral instincts often lead us to automatic, non-reflective moral judgments. These judgments are in many cases automatic and emotionally charged, and not the result of a conscious and deliberate evaluation. Haidt (2001), for example, conducted an experiment in which subjects were presented the story of a brother and sister who decided to have sex (enjoying it without remorse, making sure that there would be no procreation and keeping it secret). It was a unanimous opinion that the behavior was morally wrong, but people had serious difficulties to argue it: either irrelevant reasons were given (such as the community would feel offended, something impossible as the story clarifies that the sexual encounter was kept secret), or the inability to express the rejection was expressed. This led Haidt to argue that rather than moral reasoning, people make a moral rationalization: unfortunately for TOR, the judgment precedes the reasons.

This interpretation leads to a particularly problematic issue for TOR as a whole, and not just for its explanation of normative beliefs: the pervasive nature of rationalization. Gazzaniga (2011) has provided a solid set of experimental evidence for the existence of a process of the left hemisphere of our brain that he calls "the interpreter", which is responsible for developing coherent *post hoc* explanations of actions and emotions. Neuroscience has indicated that this interpreter "plays" with the perception of time and our own intentions. For example, we can perceive the sequence blow-pain-escape and then explain that the pain we felt led us away, but the truth is that the actual sequence was blow-escape-pain. The interpreter "cannot stand" the idea of our action as caused by something other than desires and beliefs, and our beliefs as being caused by something other than reasons. In the field of moral judgments, the actual and the perceived sequence may not match, the moral emotion preceding the reasons for it, which actually are formulated *a posteriori*.

Unfortunately for TOR, neuroscience is providing strong evidence for this non-reflective, automatic character of moral judgments. Faced with different types of moral dilemmas, individuals experience a conflict between brain areas responsible for emotion (which would be triggered as a result of our moral

intuitions) and areas responsible for logical reasoning. In cases in which the involvement of the subject is colder or distant (like pushing a button involving the death of a person and the salvation of five), the latter take control. In cases where the involvement is more direct (like killing someone with your bare hands to save the lives of five others), the former take control (Greene et al., 2001). This has been confirmed by Koenigs et al. (2007), who show the role of the emotional brain in moral judgments by studying patients with localized brain damage in those areas. Emotional drives, therefore, have a crucial role in much of our moral judgments (Nichols, 2004). In general, these judgments result from an interaction between emotion and cognition (Jeurissen et al., 2014), but some triggers lead to the dominance of one or the other process (see, for example, Hristova et al., 2014).

What can TOR (a theory that aspires to be a *general* theory of behavior) say about these universal moral judgments, their automatic and not reflective character, their dependence on the emotional brain, the blocking of the rational brain and the subsequent rationalizations, the continuity of some moral traits with other primates, and their ontogenetically early appearance? I am afraid it cannot say anything.

An example: altruistic punishment. Some moral judgments mobilized in certain economic decisions also seem to be automatic, visceral. For example, those taking place in the ultimatum game. This game is an experimental design in which two individuals interact anonymously. An individual A (the proposer) has to make a proposition on how to share a certain amount (say € 100) with the individual B (the responder). If B accepts the proposal, the division becomes effective, but if B rejects, both will leave empty handed. The RCT prediction is clear: the proposer will offer to keep € 99 and transfer € 1, and the responder will accept the offer because it would be irrational to reject a positive amount. However, the experimental results show that most of the proposals are between 40% and 50% of the amount to be distributed, and propositions below 20% have a 0.4-0.6 probability of being rejected (Fehr and Schmidt, 2006: 622).

In several texts, Boudon stated that the generally fair proposals are evidence against RCT (see, for example, 1998b:180; 2006:156; 2009:49 and 90), and they certainly are. But Boudon neglects that the behavior of the respondent is equally relevant. Respondents facing unequal proposals usually reject them. Thus, the responder assumes a cost (he waves a positive amount) in what is obviously a punishment to the proposer for his inequity. This behavior has been called "*altruistic punishment*". Although to our knowledge Boudon did not address the interpretation of this behavior, it seems clear that it would be interpreted from TOR as a behavior based on normative reasons such as "X is unfair". Although the proposer is anonymous and the interaction is one-shot, Boudon could argue that the respondent punishes someone who violates a moral principle as equity because the subject observes a moral principle consistent with punishing those who violate a moral principle such as equity, but the explanation here would become circular.

In any case, the stab to TOR comes from the genetic and neurobiological studies showing that there is something more than normative reasons behind altruistic punishment. Boudon knew the experimental results suggesting that the dorsolateral frontal cortex plays a role in altruistic punishment (2009: 112), so if that area is neutralized in a subject receiving a very unequal proposal, the subject still judges the proposal as unfair but he accepts it (i.e., he does not punish the proposer) (Koenigs and Tranel, 2007; van't Wout, 2005). Boudon's conclusion is that RCT is only applicable in those cases when a part of the brain is neutralized. Surprisingly he did not notice that this result also has implications for TOR, since the normal functioning of the dorsolateral frontal cortex (an area of the prefrontal cortex) does not move the subject away from a decision based on instrumental reasons, but away from a decision based on reasons (in general). In that line, and using brain scanning techniques, Sanfey (2004) has shown that rejections are based on visceral disgust: unequal proposals make us feel bad. Against this emotional rejection, punishment feels like a good compensation to us: also by brain scanning, DeQuervain et al. (2004) have shown that people get pleasure from the punishment of norm violators. In other words, we punish because of the negative feelings we experience when we are victims of injustice and because by punishing we experience the pleasure necessary to compensate for those feelings. In fact, subjects reject unequal proposals even when it leads to greater inequality, so that the goal of restoring equity could not explain the punishment (Moll and OliveiraSouza, 2007). As occurs in other situations of moral decision, guts act before reasons. The "moral reasoning" appears *a posteriori* to justify a behavior driven by forces that are not reasons.

Studies on the role of hormones also suggest that altruistic punishment is not driven exclusively by reasons. Research has shown that the probability of rejecting an unfair offer is greater in those with low levels of platelet serotonin (Emanuele et al., 2008) and among men with high testosterone levels (Burnham, 2007). Regarding testosterone, it appears to involve a reduction in the activity of the orbitofrontal cortex (another area of the prefrontal cortex), a brain region responsible for self-regulation and impulse control (Mehta and Beer, 2010), which comes to confirm the visceral, reactive character of this behavior. To the extent that genes play a role in encoding receptors of hormones that are relevant to behavior, genetic studies are also contributing significant evidence. For example, the dopamine D4 receptor (DRD4) gene appears to have a role in the rejection of offers in the ultimatum game (Zhong et al., 2010). In the field of genetics, but by means of twin studies, the heritability of these responses has been estimated at 42% (Wallace et al., 2007).

And again, what can TOR say about the automaticity of these behaviors, their relation to the functioning of certain brain areas, their connection with other hormones and genes responsible for those hormones' receptors, and their relative heritability? Can a theory that aspires to offer final explanations and to be a general theory of behavior ignore all these influences? As a result of these

findings, TOR is negatively affected in several of the criteria for evaluating theories: (1) in the *acceptability of the propositions*, since it establishes reasons as the only causes when empirical evidence points in another direction; (2) in *external logical consistency* and *acceptability of the assumptions*, since it assumes that a capacity for general reasoning applies equally in the formation of positive and normative beliefs when the evidence points to distinguishable cognitive processes; and (3) in *explanatory scope*, since it has nothing to say about the presented findings.

Faced with these problems of TOR, EP is proving to be a fertile matrix from where these evidences arise or to which they can be integrated. From an evolutionary point of view, the role of altruistic punishment appears to be twofold: first, it aims to increase the levels of cooperation (thus making available the adaptive advantages thereof), as the awareness of the existence of this type of behavior can deter defection (Fehr and Gatcher, 2002; Yamagishi, 1986); and second, it reduces the initial adaptive advantage of free-riders (Price, Cosmides and Tooby, 2002). Once this feature has been selected for its adaptive functions, it would remain in us as an instinct, that is, a precoded action tendency (punishing the free-rider) triggered by a disgust/seeking pleasure emotion that has been programmed to be triggered by certain cues (e.g., intentions assessed as hostile or unfair), then taking control of our behavioral reaction. By posing altruistic punishment as a cognitive instinct, the findings of genetic and neurobiological studies are easily integrated into a single framework (when they are not directly inferred from it). Thus, EP is safe from the problems that TOR suffered with several of the theoretical evaluation criteria.

5. Which *general* theory of behavior?

As mentioned above, a general theory of behavior is one that, because of its strength, is entitled to be employed as a "by default" theory in the explanation. The empirical findings of the behavioral sciences that we have discussed so far imply a negative reassessment of the strength of TOR, and especially a rejection of the necessarily final nature of its explanations. According to the idea of "stopping rules" in the microfoundation of social phenomena, sociologists could stop at the level of reasons without accounting for the neurophysiological processes that support them. In my opinion, this is right for many cases of sociological explanation. In these cases, good reasons appear to cause the behavior of individuals. However, the empirical evidence presented in this paper supports the view that in many other occasions, either reasons are systematically biased by biological causes, or these causes cause behavior, thus reasons are mere rationalizations. In either case, a reason-based explanation would be insufficient, and in some of them, wrong. TOR cannot claim the right to be used as a "by default" theory in the explanation of social behavior.

In the comparative evaluation we have made between TOR and EP, the latter is shown to be clearly superior. Can EP appeal to that strength

to stand as a *general* theory of behavior? In short: today probably it cannot, but it could do so in the future. The key will be its ability to accommodate the reason-based explanation in its framework. If EP is able to provide an interpretive framework that clarifies the conditions required for triggering a deliberative route and those required for triggering a more automatic, heuristic route, then choosing this framework "by default" in explaining social behavior would be the least bad alternative. The behavioral sciences of this century will have to work on building models that integrate reasons and biological causes. The evolutionary framework is a serious candidate to do the job.

Bibliographic references

Anokhin, A. P.; Anokhin, S.; Grant, J. D. and Heath, A. C. (2011). "Heritability of delay discounting in adolescence: a longitudinal twin study". *Behavior genetics*, 41 (2), 175-183. <http://dx.doi.org/10.1007/s10519-010-9384-7>

Baillargeon, R. (1987). "Object permanence in 3½-and 4½-month-old infants". *Developmental psychology*, 23 (5), 655. <http://dx.doi.org/10.1037/0012-1649.23.5.655>

Barrett, H. C. (2005). "Adaptations to Predators and Prey". In: Buss, D. M. (ed.). *The handbook of evolutionary psychology*. John Wiley & Sons.

Baumgartner, T.; Heinrichs, M.; Vonlanthen, A.; Fischbacher, U. and Fehr, E. (2008). "Oxytocin shapes the neural circuitry of trust and trust adaptation in humans". *Neuron*, 58 (4), 639-650. <http://dx.doi.org/10.1016/j.neuron.2008.04.009>

Berg, J.; Dickhaut, J.; McCabe, K. (1995). "Trust, Reciprocity, and Social History". *Games and Economic Behavior*, 10, 122-142. <http://dx.doi.org/10.1006/game.1995.1027>

Boudon, R. (1989). "Subjective rationality and the explanation of behavior". *Rationality and Society*, 1 (2), 173-196. <http://dx.doi.org/10.1177/1043463189001002002>

— (1990). *The art of self-persuasion. The social explanation of false beliefs*. Malden: Polity Press.

— (1996). "The 'Cognitivist Model'. A generalized 'Rational Choice Model'". *Rationality and Society*, 8 (2), 123-150. <http://dx.doi.org/10.1177/104346396008002001>

— (1997). "The moral sense». *International Sociology*, 12 (1), 5-24. <http://dx.doi.org/10.1177/026858097012001001>

— (1998a). "Limitations of Rational Choice Theory". *American Journal of Sociology*, 104 (3), 817-828.

— (1998b). "Social mechanisms without black boxes". In: Hedström, P. and Swedberg, R. (eds.). (1998). *Social mechanisms: An analytical approach to social theory*. Cambridge University Press.

— (2000). "Reasons, cognition and society". *Mind & Society*, 1, 41-56. <http://dx.doi.org/10.1007/BF02512228>

— (2001). *The origin of values. Sociology and philosophy of beliefs*, New Jersey: Transactions Publishers.

— (2002). "Sociology that Really Matters". *European Sociological Review*, 18 (3), 371-378.
<http://dx.doi.org/10.1093/esr/18.3.371>
— (2003). "Beyond Rational Choice Theory". *Annual Review of Sociology*, 29, 1-21.
<http://dx.doi.org/10.1146/annurev.soc.29.010202.100213>
— (2006). "*Homo sociologicus*: neither a rational not an irrational idiot". *Papers. Revista de Sociologia*, 80, 149-169.
— (2007) "¿Qué teoría del comportamiento para las ciencias sociales?". *Revista Española de Sociología*, 8, 5-21.
— (2009). *La racionalidad en las ciencias sociales.* Madrid: Nueva Visión.
— (2012). "'Analytical Sociology' and the explanation of beliefs". *Revue Européenne des Sciences Sociales*, 50-2, 7-34.
Brosnan, S. F.; Jones, O. D.; Lambeth, S. P.; Mareno, M. C.; Richardson, A. S. and Schapiro, S. J. (2007). "Endowment effects in chimpanzees". *Current Biology*, 17 (19), 1704-1707.
<http://dx.doi.org/10.1016/j.cub.2007.08.059>
Brown, D. E. (1991). *Human universals*. Philadelphia: Temple University Press.
Burnham, T. C. (2007). "High-testosterone men reject low ultimatum game offers". *Proceedings of the Royal Society B: Biological Sciences*, 274 (1623), 2327-2330.
<http://dx.doi.org/10.1098/rspb.2007.0546>
Buss, D. M. (2000). *The dangerous passion: Why jealousy is as necessary as love and sex*. Free Press.
Buss, D. M. *et al.* [and 50 additional authors]. (1990). "International preferences in selecting mates: A study of 37 societies". *Journal of Cross Cultural Psychology*, 21, 5-47.
<http://dx.doi.org/10.1177/0022022190211001>
Carpenter, J. P.; García, J. R. and Lum, J. K. (2011). "Dopamine receptor genes predict risk preferences, time preferences, and related economic choices". *Journal of Risk and Uncertainty*, 42 (3), 233-261.
<http://dx.doi.org/10.1007/s11166-011-9115-3>
Chen, M. K., Lakshminarayanan, V. and Santos, L. R. (2006). "How basic are behavioral biases? Evidence from capuchin monkey trading behavior". *Journal of Political Economy*, 114 (3), 517-537.
Connellan, J.; Baron-Cohen, S.; Wheelwright, S.; Batki, A. and Ahluwalia, J. (2000). "Sex differences in human neonatal social perception". *Infant Behavior and Development*, 23 (1), 113-118.
<http://dx.doi.org/10.1016/S0163-6383(00)00032-1>
Cosmides, L. and Tooby, J. (1992). "Cognitive Adaptations for Social Exchange". In: Barkow, J; Cosmides, L.; Tooby, J. (eds.), *The adapted mind: Evolutionary psychology and the generation of culture.* New York: Oxford University Press, 163-228.
— (1994). "Origins of domain-specificity: The evolution of functional organization". In: Hirschfeld, L. and Gelman, S. (eds.), *Mapping the Mind: Domain-specificity in cognition and culture.* New York: Cambridge University Press.
— (2005). "Neurocognitive adaptations designed for social exchange". In: Buss, D. (ed.) *The handbook of evolutionary psychology*, Wiley, 584-627.
Darwin, C. (1871). *The descent of man*. D. Appleton and Company.
DeBruine, L. M. (2005). "Trustworthy but not lust-worthy: Context-specific effects of facial resemblance". *Proceedings of the Royal Society B: Biological Sciences*, 272 (1566), 919-922.
<http://dx.doi.org/10.1098/rspb.2004.3003>

De Quervain, D. J. F.; Fischbacher, U.; Treyer, V.; Schellhammer, M.; Schnyder, U.; Buck, A. and Fehr, E. (2004). "The neural basis of altruistic punishment". *Science* 27 August 2004: 305 (5688), 1254-1258. <http://dx.doi.org/10.1126/science.1100735>

De Martino, B.; Camerer, C. F. and Adolphs, R. (2010). "Amygdala damage eliminates monetary loss aversion". *Proceedings of the National Academy of Sciences*, 107 (8), 3788-3792. <http://dx.doi.org/10.1073/pnas.0910230107>

De Waal, F. B. M. (1996). *Good natured: The origins of right and wrong in humans and other animals*. Cambridge: Harvard University Press.

Dreber, A.; Apicella, C. L.; Eisenberg, D. T.; García, J. R.; Zamore, R. S.; Lum, J. K. and Campbell, B. (2009). "The 7R polymorphism in the dopamine receptor D_4 gene (*DRD4*) is associated with financial risk taking in men". *Evolution and Human Behavior*, 30(2), 85-92. <http://dx.doi.org/10.1016/j.evolhumbehav.2008.11.001>

Dunbar, R; Barrett, L. and Lycett, J. (2007). *Evolutionary psychology. A beginner's guide*. Oxford: Oneworld Publications.

Dutton, D. (2009). *The art instinct: beauty, pleasure, & human evolution*. Oxford: Oxford University Press.

Elster, J. (2007). *Explaining social behavior: more nuts and bolts for the social science*. New York: Cambridge University Press.

Emanuele, E.; Brondino, N.; Bertona, M.; Re, S. and Geroldi, D. (2008). "Relationship between platelet serotonin content and rejections of unfair offers in the ultimatum game". *Neuroscience Letters*, 437 (2), 158-161. <http://dx.doi.org/10.1016/j.neulet.2008.04.006>

Engell, A. D.; Haxby, J. V. and Todorov, A. (2007). "Implicit trustworthiness decisions: automatic coding of face properties in the human amygdale". *Journal of Cognitive Neuroscience*, 19 (9), 1508-1519. <http://dx.doi.org/10.1162/jocn.2007.19.9.1508>

Fehr, E.; Fischbacher, U. and Kosfeld, M. (2005). "Neuroeconomic foundations of trust and social preferences: initial evidence". *American Economic Review*, 346-351.

Fehr, E. and Gächter, S. (2002). "Altruistic punishment in humans". *Nature*, 415 (6868), 137-140. <http://dx.doi.org/10.1038/415137a>

Fehr, E. and Schmidt, K. M. (2006) "The economics of fairness, reciprocity and altruism: experimental evidence and new theories". In: Kolm, S. C. and Ythier, J. M. (eds.) *Handbook of the Economics of Giving, Altruism and Reciprocity.* Vol. 1. Amsterdam: Elsevier.

Garbarino, E.; Slonim, R. and Sydnor, J. (2011). "Digit ratios (2D: 4D) as predictors of risky decision making for both sexes". *Journal of Risk and Uncertainty*, 42 (1), 1-26. <http://dx.doi.org/10.1007/s11166-010-9109-6>

Gazzaniga, M. S. (2011). ¿Quién manda aquí? El libre albedrío y la ciencia del cerebro. Barcelona: Paidós.

Gigerenzer, G. (2002). *Calculated risks: how to know when numbers deceive you*. New York: Simon & Schuster.

Gettler, L. T.; McDade, T. W.; Feranil, A. B. and Kuzawa, C. W. (2011). "Longitudinal evidence that fatherhood decreases testosterone in human males". *Proceedings of the National Academy of Sciences*, 108 (39), 16194-16199. <http://dx.doi.org/10.1073/pnas.1105403108>

Gettler, L. T.; McKenna, J. J.; McDade, T. W.; Agustin, S. S. and Kuzawa, C. W. (2012). "Does cosleeping contribute to lower testosterone levels in fathers? Evidence from the Philippines". *PloS one*, 7 (9), e41559. <http://dx.doi.org/10.1371/journal.pone.0041559>

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. and Cohen, J. D. (2001). "An fMRI investigation of emotional engagement in moral judgment". *Science*, 293 (5537), 2105-2108. <http://dx.doi.org/10.1126/science.1062872>

Haidt, J. (2001). "The emotional dog and its rational tail: a social intuitionist approach to moral judgment". *Psychological Review*, 108 (4), 814. <http://dx.doi.org/10.1037/0033-295X.108.4.814>

Haidt, J. and Graham, J. (2007a). "Planet of the Durkheimians, where community, authority, and sacredness are foundations of morality". In: Jost; J., Kay; A. C. and Thorisdottir, H. (eds.), *Social and psychological bases of ideology and system justification,* 371-401. New York: Oxford University Press.

— (2007b) "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize". *Social Justice Research*, 20 (1), 98-116. <http://dx.doi.org/10.1007/s11211-007-0034-z>

Haidt, J. and Kesebir, S. (2010). "Morality". In: Fiske, S. T.; Gilbert, D. T. and Lindzey, G. (eds.). (2010). *Handbook of social psychology*, Wiley.

Haidt, J. and Joseph, C. (2004). "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues". *Daedalus*, 55-66. <http://dx.doi.org/10.1162/0011526042365555>

Hariri, A. R.; Brown, S. M.; Williamson, D. E.; Flory, J. D.; De Wit, H. and Manuck, S. B. (2006). "Preference for immediate over delayed rewards is associated with magnitude of ventral striatal activity". *The Journal of Neuroscience*, 26 (51), 13213-13217. <http://dx.doi.org/10.1523/JNEUROSCI.3446-06.2006>

Hristova, E.; Kadreva, V. and Grinberg, M. (2014). "Emotions and Moral Judgment: A Multimodal Analysis". In: Bassis, Simone; Esposito, Anna; Morabito and Francesco Carlo (eds.) *Recent Advances of Neural Network Models and Applications,* 413-421. New York: Springer International Publishing.

He, Q.; Xue, G.; Chen, C. *et al.* (2010). "Serotonin transporter gene-linked polymorphic region (5-HTTLPR) influences decision making under ambiguity and risk in a large Chinese sample". *Neuropharmacology*, 59 (6), 518-526. <http://dx.doi.org/10.1016/j.neuropharm.2010.07.008>

Hwang, J.; Kim, S. and Lee, D. (2009). "Temporal discounting and inter-temporal choice in rhesus monkeys". *Frontiers in behavioral neuroscience*, 3, 9, 1-13. <http://dx.doi.org/10.3389/neuro.08.009.2009>

Jeurissen, D.; Sack, A. T.; Roebroeck, A.; Russ, B. E. and Pascual-Leone, A. (2014). "TMS Affects Moral Judgment, Showing the Role of DLPFC and TPJ in Cognitive and Emotional Processing". *Frontiers in Neuroscience*, 8, 18. <http://dx.doi.org/10.3389/fnins.2014.00018>

Kahneman, D. and Tversky, A. (1984). "Choices, values, and frames". *American Psychologist*, 39 (4), 341-350. <http://dx.doi.org/10.1037/0003-066X.39.4.341>

Kahneman, D.; Knetsch, J. L. and Thaler, R. H. (1990). "Experimental Tests of the Endowment Effect and the Coase Theorem". *Journal of Political Economy*, 98 (6), 1325-1348.

Kayser, A. S.; Allen, D. C.; Navarro-Cebrian, A.; Mitchell, J. M. and Fields, H. L. (2012). "Dopamine, corticostriatal connectivity, and intertemporal choice". *The Journal of Neuroscience*, 32 (27), 9402-9409. <http://dx.doi.org/10.1523/JNEUROSCI.1180-12.2012>

Kéri, S. and Kiss, I. (2011). "Oxytocin response in a trust game and habituation of arousal". *Physiology & Behavior*, 102 (2), 221-224. <http://dx.doi.org/10.1016/j.physbeh.2010.11.011>

Knetsch, J. L. (1989). "The endowment effect and evidence of nonreversible indifference curves". *The American Economic Review*, 79 (5), 1277-1284.

Koenigs, M.; Young, L.; Adolphs, R.; Tranel, D.; Cushman, F.; Hauser, M. and Damasio, A. (2007). "Damage to the prefrontal cortex increases utilitarian moral judgements". *Nature*, 446 (7138), 908-911. <http://dx.doi.org/10.1038/nature05631>

Koenigs, M. and Tranel, D. (2007). "Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game". *The Journal of Neuroscience*, 27 (4), 951-956. <http://dx.doi.org/10.1523/JNEUROSCI.4606-06.2007>

Kosfeld, M.; Heinrichs, M.; Zak, P. J.; Fischbacher, U. and Fehr, E., (2005). "Oxytocin increases trust in humans". *Nature*, 435, 673-676. <http://dx.doi.org/10.1038/nature03701>

Krueger, F.; Parasuraman, R.; Iyengar, V.; Thornburg, M.; Weel, J.; Lin, M. and Lipsky, R. H. (2012). "Oxytocin receptor genetic variation promotes human trust behavior". *Frontiers in Human Neuroscience*, 6. <http://dx.doi.org/10.3389/fnhum.2012.00004>

Kuhnen, C. M. and Chiao, J. Y. (2009). "Genetic Determinants of Financial Risk Taking". *PLoS ONE* 4(2) e4362. <http://dx.doi.org/10.1371/journal.pone.0004362>

Levy, M. S. (2010). "Evolution of Risk Aversion: The "Having Descendants Forever" Approach". Working Paper.

Li, Y. J.; Kenrick, D. T.; Griskevicius, V. and Neuberg, S. L. (2012). "Economic decision biases and fundamental motivations: How mating and self-protection alter loss aversion". *Journal of Personality and Social Psychology*, 102 (3), 550. <http://dx.doi.org/10.1037/a0025844>

Litt, A.; Eliasmith, C. and Thagard, P. (2006). "Why losses loom larger than gains: Modeling neural mechanisms of cognitive-affective interaction". *Proceedings of the twenty-eighth annual meeting of the Cognitive Science Society*, 495-500.

Lizón, A. (2010). «Encrucijadas teóricas en la sociología del siglo xx». *Papers. Revista de Sociologia*, 95 (2), 389-420

Machery, E. (forthcoming). "Discovery and confirmation in evolutionary psychology". In: Prinz, Jesse J. (ed.). *The Oxford Handbook of Philosophy of Psychology*. Oxford University Press.

Manuck, S. B.; Flory, J. D.; Muldoon, M. F. and Ferrell, R. E. (2003). "A neurobiology of intertemporal choice". In: Loewenstein, G.; Read, D. and Baumeister, R. F. (eds.) *Time and decision: Economic and psychological perspectives on intertemporal choice*. New York: Russell Sage Foundation, 139-172.

McDermott, R.; Fowler, J. H. and Smirnov, O. (2008). "On the evolutionary origin of prospect theory preferences". *Journal of Politics*, 70 (2), 335-50. <http://dx.doi.org/10.1017/S0022381608080341>

Mehta, P. H. and Beer, J. (2010). "Neural mechanisms of the testosterone–aggression relation: The role of orbitofrontal cortex". *Journal of Cognitive Neuroscience*, 22 (10), 2357-2368.
<http://dx.doi.org/10.1162/jocn.2009.21389>

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. Oxford University Press.

Moll, J. and De Oliveira-Souza, R. (2007). "Moral judgments, emotions and the utilitarian brain". *Trends in cognitive sciences*, 11 (8), 319-321.
<http://dx.doi.org/10.1016/j.tics.2007.06.001>

Moon, C.; Cooper, R. P. and Fifer, W. P. (1993). "Two-day-olds prefer their native language". *Infant behavior and development*, 16 (4), 495-500.
<http://dx.doi.org/10.1016/0163-6383(93)80007-U>

Murphy, S. E.; Longhitano, C.; Ayres, R. E.; Cowen, P. J.; Harmer, C. J. and Rogers, R. D. (2009). "The role of serotonin in nonnormative risky choice: the effects of tryptophan supplements on the 'reflection effect' in healthy adult volunteers". *Journal of Cognitive Neuroscience*, 21 (9), 1709-1719.
<http://dx.doi.org/10.1162/jocn.2009.21122>

Orians, G. H. and Heerwagen, J. H. (1992). "Evolved responses to landscapes». In: Barkow, J. H.; Cosmides, L. E.; Tooby, J. E. *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford University Press.

Peters, J. and Büchel, C. (2011). "The neural mechanisms of inter-temporal decision-making: understanding variability". *Trends in cognitive sciences*, 15 (5), 227-239.
<http://dx.doi.org/10.1016/j.tics.2011.03.002>

Pinker, S. (2003). *The blank slate: The modern denial of human nature*. Londres: Penguin Books.

— (2008). "The moral instinct", *The New York Times Magazine*, 13 January 2008.

Price, M. E.; Cosmides, L. and Tooby, J. (2002). "Punitive sentiment as an anti-free rider psychological device". *Evolution and Human Behavior*, 23 (3), 203-231.
<http://dx.doi.org/10.1016/S1090-5138(01)00093-9>

Rakison, D. H. and Derringer, J. (2008). "Do infants possess an evolved spider-detection mechanism?". *Cognition*, 107 (1), 381-393.
<http://dx.doi.org/10.1016/j.cognition.2007.07.022>

Rodrigues S. M.; Saslow, L. R.; García, N.; John O. P. and Keltner, D. (2009). "Oxytocin receptor genetic variation relates to empathy and stress reactivity in humans". *Proc. Natl. Acad. Sci. U.S.A.* 106 (50): 21437-21441.
<http://dx.doi.org/10.1073/pnas.0909579106>

Sanfey, A. G. (2004). "Neural computations of decision utility". *Trends in Cognitive Sciences*, 8 (12), 519-521.
<http://dx.doi.org/10.1016/j.tics.2004.10.006>

Saphire-Bernstein, S.; Way, B. M.; Kim, H. S.; Sherman, D. K. and Taylor, S. E. (2011). "Oxytocin receptor gene (OXTR) is related to psychological resources". *Proc. Natl. Acad. Sci. U.S.A.,* 108 (37), 15118–22.
<http://dx.doi.org/10.1073/pnas.1113137108>

Schmitt, D. P. and Pilcher, J. J. (2004). "Evaluating evidence of psychological adaptation: How do we know one when we see one?". *Psychological Science*, 15 (10), 643-649.
<http://dx.doi.org/10.1111/j.0956-7976.2004.00734.x>

Shiv, B.; Loewenstein, G.; Bechara, A.; Damasio, H. and Damasio, A. R. (2005). "Investment behavior and the negative side of emotion". *Psychological Science*, 16 (6), 435-439.

Simon, H. (1982). *Models of bounded rationality*. Cambridge: MITT Press.

Sozou, P. (1998). "On hyperbolic discounting and uncertain hazard rates". *Proceedings of the Royal Society B: Biological Sciences*, 265 (1409), 2015-2020. <http://dx.doi.org/10.1098/rspb.1998.0534>

Spelke, E. S. (1990). "Principles of object perception". *Cognitive Science*, 14, 29-56. <http://dx.doi.org/10.1207/s15516709cog1401_3>

Stanton, S. J.; O'Dhaniel, A.; McLaurin, R. E.; Kuhn, C. M.; LaBar, K. S.; Platt, M. L. and Huettel, S. A. (2011). "Low-and high-testosterone individuals exhibit decreased aversion to economic risk". *Psychological Science*, 22 (4), 447-453. <http://dx.doi.org/10.1177/0956797611401752>

Tabak, B. A.; McCullough, M. E.; Carver, C. S.; Pedersen, E. J. and Cuccaro, M. L. (2013). "Variation in oxytocin receptor gene (OXTR) polymorphisms is associated with emotional and behavioral reactions to betrayal". *Social Cognitive and Affective Neuroscience*, nst042. <http://dx.doi.org/10.1093/scan/nst042>

Takahashi, C.; Yamagishi, T.; Tanida, S.; Kiyonari, T. and Kanazawa, S. (2006). "Attractiveness and cooperation in social exchange". *Evolutionary Psychology,* 4, 315-329.

Todorov, A.; Baron, S. G. and Oosterhof, N. N. (2008). "Evaluating face trustworthiness: a model based approach". *Social Cognitive and Affective Neuroscience*, 3 (2), 119-127. <http://dx.doi.org/10.1093/scan/nsn009>

Tomasello, M. (2009). *Why we cooperate*. Cambridge: MIT Press.

Tooby, J. and Cosmides, L. (1992). "The psychological foundations of culture". In: Barkow, J.; Cosmides, L. and Tooby, J. (eds.), *The adapted mind: Evolutionary psychology and the generation of culture.* New York: Oxford University Press.

Tost, H.; Kolachana, B.; Hakimi, S.; Lemaitre, H.; Verchinski, B. A.; Mattay, V. S.; Weinberger, D. R. and Meyer-Lindenberg, A. (2010). "A common allele in the oxytocin receptor gene (OXTR) impacts prosocial temperament and human hypothalamic-limbic structure and function". *Proc. Natl. Acad. Sci. U.S.A.,* 107 (31): 13936-13941. <http://dx.doi.org/10.1073/pnas.1003296107>

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge University Press.

Tversky, A. and Kahneman, D. (1981). "The framing of decisions and the psychology of choice". *Science*, 211, 4481, 453-458. <http://dx.doi.org/10.1126/science.7455683>

Van Ijzendoorn, M. H. and Bakermans-Kranenburg, M. J. (2012). "A sniff of trust: meta-analysis of the effects of intranasal oxytocin administration on face recognition, trust to in-group, and trust to out-group". *Psychoneuroendocrinology*, 37 (3), 438-443. <http://dx.doi.org/10.1016/j.psyneuen.2011.07.008>

Van't Wout, M.; Kahn, R. S.; Sanfey, A. G. and Aleman, A. (2005). "Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making". *Neuroreport*, 16 (16), 1849-1852.

Wallace, B.; Cesarini, D.; Lichtenstein, P. and Johannesson, M. (2007). "Heritability of ultimatum game responder behavior". *Proceedings of the National Academy of Sciences of the United States of America,* 104, 15631-15634. <http://dx.doi.org/10.1073/pnas.0706642104>

Walum, H.; Lichtenstein, P.; Neiderhiser, J. M.; Reiss, D.; Ganiban, J. M.; Spotts, E. L. and Westberg, L. (2012). "Variation in the oxytocin receptor gene is associated with pair-bonding and social behavior". *Biological Psychiatry*, 71 (5), 419-426. <http://dx.doi.org/10.1016/j.biopsych.2011.09.002>

Winston, J. S.; Strange, B. A.; O'Doherty, J. and Dolan, R. J. (2002). "Automatic and intentional brain responses during evaluation of trustworthiness of faces". *Nature Neuroscience*, 5 (3), 277-283. <http://dx.doi.org/10.1038/nn816>

Yamagishi, T. (1986). "The provision of a sanctioning system as a public good". *Journal of Personality and Social Psychology*, 51 (1), 110. <http://dx.doi.org/10.1037/0022-3514.51.1.110>

Zak, P. J.; Kurzban, R. and Matzner, W. T. (2005). "Oxytocin is associated with human trustworthiness". *Hormones and Behavior*, 48 (5), 522-527. <http://dx.doi.org/10.1016/j.yhbeh.2005.07.009>

Zhong, S.; Israel, S.; Xue, H.; Ebstein, R. P. and Chew, S. H. (2009). "Monoamine oxidase A gene (MAOA) associated with attitude towards longshot risks". *PLoS One*, 4 (12), e8516. <http://dx.doi.org/10.1371/journal.pone.0008516>

Zhong, S.; Israel, S.; Shalev, I.; Xue, H.; Ebstein, R. P. and Chew, S. H. (2010). "Dopamine D4 receptor gene associated with fairness preference in ultimatum game". PLoS One, 5 (11), e13765. <http://dx.doi.org/10.1371/journal.pone.0013765>

Zyphur, M. J.; Narayanan, J.; Arvey, R. D. and Alexander, G. J. (2009). "The genetics of economic risk preferences". *Journal of Behavioral Decision Making*, 22 (4), 367-377. <http://dx.doi.org/10.1002/bdm.643>